PHASE-CHANGE TECHNOLOGY AND THE FUTURE OF MAIN MEMORY

PHASE-CHANGE MEMORY MAY ENABLE CONTINUED SCALING OF MAIN MEMORIES, BUT PCM HAS HIGHER ACCESS LATENCIES, INCURS HIGHER POWER COSTS, AND WEARS OUT MORE QUICKLY THAN DRAM. THIS ARTICLE DISCUSSES HOW TO MITIGATE THESE LIMITATIONS THROUGH BUFFER SIZING, ROW CACHING, WRITE REDUCTION, AND WEAR LEVELING, TO MAKE PCM A VIABLE DRAM ALTERNATIVE FOR SCALABLE MAIN MEMORIES.

Benjamin C. Lee
Stanford University

Ping Zhou
Jun Yang
Youtao Zhang
Bo Zhao
University of Pittsburgh

Engin Ipek
University of Rochester

Onur Mutlu
Carnegie Mellon University

Doug Burger
Microsoft Research

Over the past few decades, memory technology scaling has provided many benefits, including increased density and capacity and reduced cost. Scaling has provided these benefits for conventional technologies, such as DRAM and flash memory, but now scaling is in jeopardy. For continued scaling, systems might need to transition from conventional *charge memory* to emerging *resistive memory*. Charge memories require discrete amounts of charge to induce a voltage, which is detected during reads. In the nonvolatile space, flash memories must precisely control the discrete charge placed on a floating gate. In volatile main memory, DRAM must not only place charge in a storage capacitor but also mitigate subthreshold charge leakage through the access device. Capacitors must be sufficiently large to store charge for reliable sensing, and transistors must be sufficiently large to exert effective control over the channel. Given these challenges, scaling DRAM beyond 40 nanometers will be increasingly difficult.[1]

In contrast, resistive memories use electrical current to induce a change in atomic structure, which impacts the resistance detected during reads. Resistive memories are amenable to scaling because they don't require precise charge placement and control. Programming mechanisms such as current injection scale with cell size. Phase-change memory (PCM), spin-torque transfer (STT) magnetoresistive RAM (MRAM), and ferroelectric RAM (FRAM) are examples of resistive memories. Of these, PCM is closest to realization and imminent deployment as a NOR flash competitor. In fact, various researchers and manufacturers have prototyped PCM arrays in the past decade.[2]

PCM provides a nonvolatile storage mechanism that is amenable to process scaling. During writes, an access transistor injects current into the storage material and thermally induces phase change, which is detected during reads. PCM, relying on analog current and thermal effects, doesn't require control over discrete electrons. As technologies scale and heating contact areas shrink, programming current scales linearly. Researchers project this PCM scaling mechanism will be more robust than that of DRAM beyond 40 nm, and it has already been demonstrated in a 32-nm device prototype.[1,3] As a scalable DRAM alternative, PCM could provide a clear road map for increasing main memory density and capacity.

Providing a path for main-memory scaling, however, will require surmounting PCM's disadvantages relative to DRAM.

0272-1732/10/$26.00 © 2010 IEEE

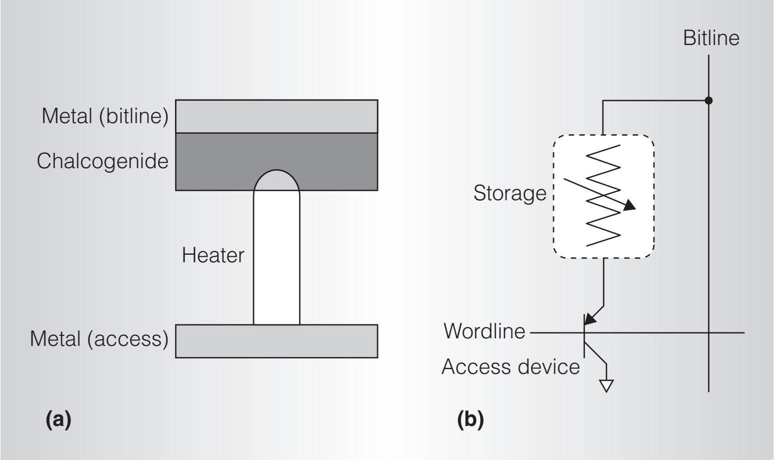

Figure 1. Storage element with heater and chalcogenide between electrodes (a), and cell structure with storage element and bipolar junction transistor (BJT) access device (b).

PCM access latencies, although only tens of nanoseconds, are still several times slower than those of DRAM. At present technology nodes, PCM writes require energy-intensive current injection. Moreover, the resulting thermal stress within the storage element degrades current-injection contacts and limits endurance to hundreds of millions of writes per cell for present process technologies. Because of these significant limitations, PCM is presently positioned mainly as a flash replacement. As a DRAM alternative, therefore, PCM must be architected for feasibility in main memory for general-purpose systems.

Today's prototype designs do not mitigate PCM latencies, energy costs, and finite endurance. This article describes a range of PCM design alternatives aimed at making PCM systems competitive with DRAM systems. These alternatives focus on four areas: row buffer design, row caching, wear reduction, and wear leveling.[2,4] Relative to DRAM, these optimizations collectively provide competitive performance, comparable energy, and feasible lifetimes, thus making PCM a viable replacement for main-memory technology.

Technology and challenges

As Figure 1 shows, the PCM storage element consists of two electrodes separated by a resistive heater and a chalcogenide (the phase-change material). $Ge_2Sb_2Te_5$ (GST) is the most commonly used chalcogenide, but others offer higher resistivity and improve the device's electrical characteristics. Nitrogen doping increases resistivity and lowers programming current, whereas GS offers lower-latency phase changes.[5,6] (GS contains the first two elements of GST, germanium and antimony, and does not include tellurium.)

Phase changes are induced by injecting current into the resistor junction and heating the chalcogenide. The current and voltage characteristics of the chalcogenide are identical regardless of its initial phase, thereby lowering programming complexity and latency.[7] The amplitude and width of the injected current pulse determine the programmed state.

PCM cells are one-transistor (1T), one-resistor (1R) devices comprising a resistive storage element and an access transistor (Figure 1). One of three devices typically controls access: a field-effect transistor (FET), a bipolar junction transistor (BJT), or a diode. In the future, FET scaling and large voltage drops across the cell will adversely affect gate-oxide reliability for unselected wordlines.[8] BJTs are faster and can scale more robustly without this vulnerability.[8,9] Diodes occupy smaller areas and potentially enable greater cell densities but require higher operating voltages.[10]

Writes

The access transistor injects current into the storage material and thermally induces phase change, which is detected during reads. The chalcogenide's resistivity captures logical data values. A high, short current pulse (*reset*) increases resistivity by abruptly discontinuing current, quickly quenching heat generation, and freezing the chalcogenide into an amorphous state. A moderate, long current pulse (*set*) reduces resistivity by ramping down current, gradually cooling the chalcogenide, and inducing crystal growth. *Set* latency, which requires longer current pulses, determines write performance. *Reset* energy, which requires higher current pulses, determines write power.

Cells that store multiple resistance levels could be implemented by leveraging intermediate states, in which the chalcogenide is partially crystalline and partially amorphous.[9,11] Smaller current slopes (slow ramp-down) produce lower resistances, and larger slopes (fast ramp-down) produce higher resistances. Varying slopes induce partial phase transitions and/or change the size and shape of the amorphous material produced at the contact area, generating resistances between those observed from fully amorphous or fully crystalline chalcogenides. The difficulty and high-latency cost of differentiating between many resistances could constrain such multilevel cells to a few bits per cell.

Wear and endurance

Writes are the primary wear mechanism in PCM. When current is injected into a volume of phase-change material, thermal expansion and contraction degrades the electrode storage contact, such that programming currents are no longer reliably injected into the cell. Because material resistivity highly depends on current injection, current variability causes resistance variability. This greater variability degrades the read window, which is the difference between programmed minimum and maximum resistances.

Write endurance, the number of writes performed before the cell cannot be programmed reliably, ranges from 10^4 to 10^9. Write endurance depends on process and differs across manufacturers. PCM will likely exhibit greater write endurance than flash memory by several orders of magnitude (for example, 10^7 to 10^8). The 2007 *International Technology Roadmap for Semiconductors* (*ITRS*) projects an improved endurance of 10^{12} writes at 32 nm.[1] Wear reduction and leveling techniques could prevent write limits from being exposed to the system during a memory's lifetime.

Reads

Before the cell is read, the bitline is precharged to the read voltage. The wordline is active-low when using a BJT access transistor (see Figure 1). If a selected cell is in a crystalline state, the bitline is discharged, with current flowing through the storage element and the access transistor. Otherwise, the cell is in an amorphous state, preventing or limiting bitline current.

Scalability

As contact area decreases with feature size, thermal resistivity increases, and the volume of phase-change material that must be melted to completely block current flow decreases. Specifically, as feature size scales down ($1/k$), contact area decreases quadratically ($1/k^2$). Reduced contact area causes resistivity to increase linearly (k), which in turn causes programming current to decrease linearly ($1/k$). These effects enable not only smaller storage elements but also smaller access devices for current injection. At the system level, scaling translates into lower memory-subsystem energy. Researchers have demonstrated this PCM scaling mechanism in a 32-nm device prototype.[1,3]

PCM characteristics

Realizing the vision of PCM as a scalable memory requires understanding and overcoming PCM's disadvantages relative to DRAM. Table 1 shows derived technology parameters from nine prototypes published in the past five years by multiple semiconductor manufacturers.[2] Access latencies of up to 150 ns are several times slower than those of DRAM. At current 90-nm technology nodes, PCM writes require energy-intensive current injection. Moreover, writes induce thermal expansion and contraction within storage elements, degrading injection contacts and limiting endurance to hundreds of millions of writes per cell at current processes. Prototypes implement $9F^2$ to $12F^2$ PCM cells using BJT access devices (where F is the feature size)—up to 50 percent larger than $6F^2$ to $8F^2$ DRAM cells.

These limitations position PCM as a replacement for flash memory; in this market, PCM properties are drastic improvements. Making PCM a viable alternative to DRAM, however, will require architecting PCM for feasibility in main memory for general-purpose systems.

Architecting a DRAM alternative

With area-neutral buffer reorganizations, Lee et al. show that PCM systems are within the competitive range of DRAM systems.[2] Effective buffering hides long PCM latencies and reduces PCM energy costs. Scalability trends further favor PCM over DRAM.

Table 1. Technology survey.

	Published prototype									
Parameter*	**Horri[6]**	**Ahn[12]**	**Bedeschi[13]**	**Oh[14]**	**Pellizer[15]**	**Chen[5]**	**Kang[16]**	**Bedeschi[9]**	**Lee[10]**	**Lee[2]**
Year	2003	2004	2004	2005	2006	2006	2006	2008	2008	**
Process, F (nm)	**	120	180	120	90	**	100	90	90	90
Array size (Mbytes)	**	64	8	64	**	**	256	256	512	**
Material	GST, N-d	GST, N-d	GST	GST	GST	GS, N-d	GST	GST	GST	GST, N-d
Cell size (μm^2)	**	0.290	0.290	**	0.097	60 nm^2	0.166	0.097	0.047	0.065 to 0.097
Cell size, F^2	**	20.1	9.0	**	12.0	**	16.6	12.0	5.8	9.0 to 12.0
Access device	**	**	BJT	FET	BJT	**	FET	BJT	Diode	BJT
Read time (ns)	**	70	48	68	**	**	62	**	55	48
Read current (μA)	**	**	40	**	**	**	**	**	**	40
Read voltage (V)	**	3.0	1.0	1.8	1.6	**	1.8	**	1.8	1.0
Read power (μW)	**	**	40	**	**	**	**	**	**	40
Read energy (pJ)	**	**	2.0	**	**	**	**	**	**	2.0
Set time (ns)	100	150	150	180	**	80	300	**	400	150
Set current (μA)	200	**	300	200	**	55	**	**	**	150
Set voltage (V)	**	**	2.0	**	**	1.25	**	**	**	1.2
Set power (μW)	**	**	300	**	**	34.4	**	**	**	90
Set energy (pJ)	**	**	45	**	**	2.8	**	**	**	13.5
Reset time (ns)	50	10	40	10	**	60	50	**	50	40
Reset current (μA)	600	600	600	600	400	90	600	300	600	300
Reset voltage (V)	**	**	2.7	**	1.8	1.6	**	1.6	**	1.6
Reset power (μW)	**	**	1620	**	**	80.4	**	**	**	480
Reset energy (pJ)	**	**	64.8	**	**	4.8	**	**	**	19.2
Write endurance (MLC)	10^7	10^9	10^6	**	10^8	10^4	**	10^5	10^5	10^8

* BJT: bipolar junction transistor; FET: field-effect transistor; GST: $Ge_2Sb_2Te_5$; MLC: multilevel cells; N-d: nitrogen doped.
** This information is not available in the publication cited.

Array architecture

PCM cells can be hierarchically organized into banks, blocks, and subblocks. Despite similarities to conventional memory array architectures, PCM has specific design issues that must be addressed. For example, PCM reads are nondestructive.

Choosing bitline sense amplifiers affects array read-access time. Voltage-based sense amplifiers are cross-coupled inverters that require differential discharging of bitline capacitances. In contrast, current-based sense amplifiers rely on current differences to create differential voltages at the amplifiers' output nodes. Current sensing is faster but requires larger circuits.[17]

In DRAM, sense amplifiers both detect and buffer data using cross-coupled inverters. In contrast, we explore PCM architectures in which sensing and buffering are separate. In such architectures, sense amplifiers drive banks of explicit latches. These latches provide greater flexibility in row buffer organization by enabling multiple buffered rows. However, they also incur area overheads. Separate sensing and buffering enables multiplexed sense amplifiers. Multiplexing also enables buffer widths narrower than array widths (defined by the total number of bitlines). Buffer width is a critical design parameter that determines the required number of expensive current-based sense amplifiers.

Buffer organizations

Another way to make PCM more competitive with DRAM is to use area-neutral buffer organizations, which have several benefits. First, area neutrality enables a competitive DRAM alternative in an industry where area and density directly impact cost and profit. Second, to mitigate fundamental PCM constraints and achieve competitive performance and energy relative to DRAM-based systems, narrow buffers reduce the number of high-energy PCM writes, and multiple rows exploit temporal locality. This locality not only improves performance, but also reduces energy by exposing additional opportunities for write coalescing. Third, as PCM technology matures, baseline PCM latencies will likely improve. Finally, process technology scaling will drive linear reductions in PCM energy.

Area neutrality. Buffer organizations achieve area neutrality through narrower buffers and additional buffer rows. The number of sense amplifiers decreases linearly with buffer width, significantly reducing area because fewer of these large circuits are required. We take advantage of these area savings by implementing multiple rows with latches far smaller than the removed sense amplifiers. Narrow widths reduce PCM write energy but negatively impact spatial locality, opportunities for write coalescing, and application performance. However, the additional buffer rows can mitigate these penalties. We examine these fundamental trade-offs by constructing area models and identifying designs that meet a DRAM-imposed area budget before optimizing delay and energy.[2]

Buffer design space. Figure 2 illustrates the delay and energy characteristics of the buffer design space for representative benchmarks from memory-intensive scientific-computing applications.[18-20] The triangles represent PCM and DRAM baselines implementing a single 2,048-byte buffer. Circles represent various buffer organizations. Open circles indicate organizations requiring less area than the DRAM baseline when using $12F^2$ cells. Closed circles indicate additional designs that become viable when considering smaller $9F^2$ cells. By default, the PCM baseline (see the triangle labeled ''PCM base'' in the figure) does not satisfy the area budget because of larger current-based sense amplifiers and explicit latches.

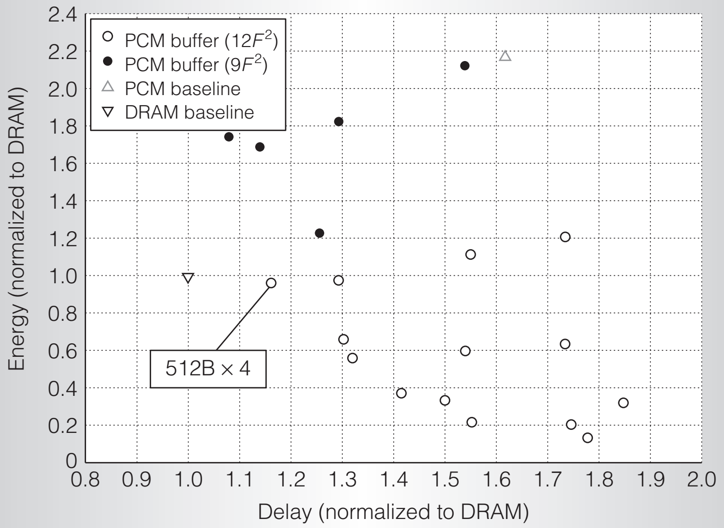

Figure 2. Phase-change memory (PCM) buffer organization, showing delay and energy averaged for benchmarks to illustrate Pareto frontier. Open circles indicate designs satisfying area constraints, assuming $12F^2$ PCM multilevel cells. Closed circles indicate additional designs satisfying area constraints, assuming smaller $9F^2$ PCM multilevel cells.

As Figure 2 shows, reorganizing a single, wide buffer into multiple, narrow buffers reduces both energy costs and delay. Pareto optima shift PCM delay and energy into the neighborhood of the DRAM baseline. Furthermore, among these Pareto optima, we observe a knee that minimizes both energy and delay: four buffers that are 512 bytes wide. Such an organization reduces the PCM delay and energy disadvantages from 1.6× and 2.2× to 1.1× and 1.0×, respectively. Although smaller $9F^2$ PCM cells provide the area for wider buffers and additional rows, the associated energy costs are not justified. In general, diminishing marginal reductions in delay suggest that area savings from $9F^2$ cells should go toward improving density, not additional buffering.

Delay and energy optimization. Using four 512-byte buffers is the most effective way to optimize average delay and energy across workloads. Figure 3 illustrates the impact of reorganized PCM buffers. Delay penalties are reduced from the original 1.60× to 1.16×. The delay impact ranges from 0.88× (swim benchmark) to 1.56× (fft

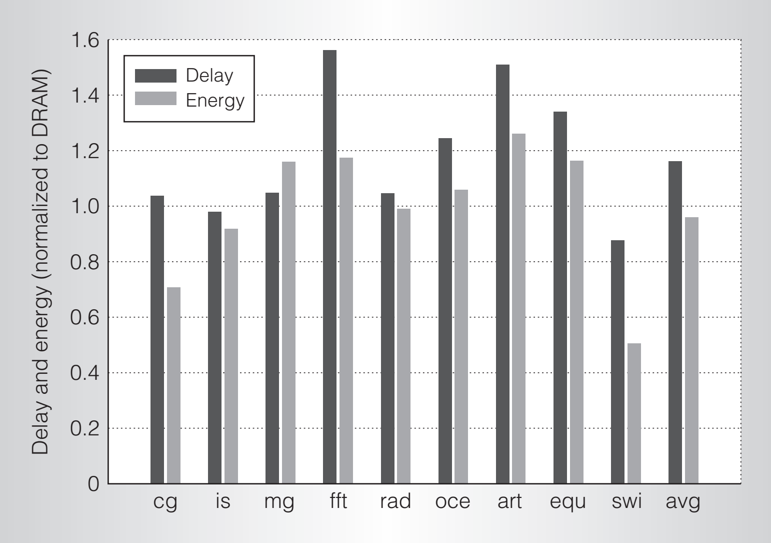

Figure 3. Application delay and energy when using PCM with optimized buffering (four 512-byte buffers) as a DRAM replacement.

benchmark) relative to a DRAM-based system. When executing benchmarks on an effectively buffered PCM, more than half of the benchmarks are within 5 percent of their DRAM performance. Benchmarks that perform less effectively exhibit low write-coalescing rates. For example, buffers can't coalesce any writes in the fft workload.

Buffering and write coalescing also reduce memory subsystem energy from a baseline of 2.2× to 1.0× parity with DRAM. Although each PCM array write requires 43.1× more energy than a DRAM array write, these energy costs are mitigated by narrow buffer widths and additional rows, which reduce the granularity of buffer evictions and expose opportunities for write coalescing, respectively.

Scaling and implications

DRAM scaling faces many significant technical challenges because scaling exposes weaknesses in both components of the one-transistor, one-capacitor cell. Capacitor scaling is constrained by the DRAM storage mechanism. Scaling makes increasingly difficult the manufacture of small capacitors that store sufficient charge for reliable sensing despite large parasitic capacitances on the bitline.

Scaling scenarios are also bleak for access transistors. As these transistors scale down, increasing subthreshold leakage makes it increasingly difficult to ensure DRAM retention times. Not only is less charge stored in the capacitor, but that charge is also stored less reliably. These trends will impact DRAM's reliability and energy efficiency in future process technologies. According to the *ITRS*, "manufacturable solutions are not known" for DRAM beyond 40 nm.[1]

In contrast, the *ITRS* projects PCM scaling mechanisms will extend to 32 nm, after which other scaling mechanisms could apply.[1] PCM scaling mechanisms have already been demonstrated at 32 nm with a novel device structure designed by Raoux et al.[3] Although both DRAM and PCM are expected to be viable at 40 nm, energy-scaling trends strongly favor PCM. Lai and Pirovano et al. have separately projected a 2.4× reduction in PCM energy from 80 nm to 40 nm.[7,8] In contrast, the *ITRS* projects that DRAM energy will fall by only 1.5× across the same technology nodes, thus reflecting the technical challenges of DRAM scaling.

Because PCM energy scales down 1.6× more quickly than DRAM energy, PCM systems will significantly outperform DRAM systems at future technology nodes. At 40 nm, PCM system energy is 61.3 percent that of DRAM, averaged across workloads. Switching from DRAM to PCM reduces energy costs by at least 22.1 percent (art benchmark) and by as much as 68.7 percent (swim benchmark). This analysis does not account for refresh energy, which could further increase DRAM energy costs. Although the *ITRS* projects constant retention time of 64 ms as DRAM scales to 40 nm,[3] less-effective access-transistor control might reduce retention times. If retention times fall, DRAM refresh energy will increase as a fraction of total DRAM energy costs.

Mitigating wear and energy

In addition to architecting PCM to offer competitive delay and energy relative to DRAM, we must also consider PCM wear mechanisms. With only 10^7 to 10^8 writes over each cell's lifetime, solutions are needed to reduce and level writes coming from the lowest-level processor cache. Zhou et al.

show that write reduction and leveling can improve PCM endurance with light circuitry overheads.[4] These schemes level wear across memory elements, remove redundant bit writes, and collectively achieve an average lifetime of 22 years. Moreover, an energy study shows PCM with low-operating power (LOP) peripheral logic is energy efficient.

Improving PCM lifetimes

An evaluation on a set of memory-intensive workloads shows that the unprotected lifetime of PCM-based main memory can last only an average of 171 days. Although Lee et al. track written cache lines and written cache words to implement partial writes and reduce wear,[2] fine-grained schemes at the bit level might be more effective. Moreover, combining wear reduction with wear leveling can address low lifetimes arising from write locality. Here, we introduce a hierarchical set of techniques that both reduce and level wear to improve the lifetime of PCM-based main memory to more than 20 years on average.

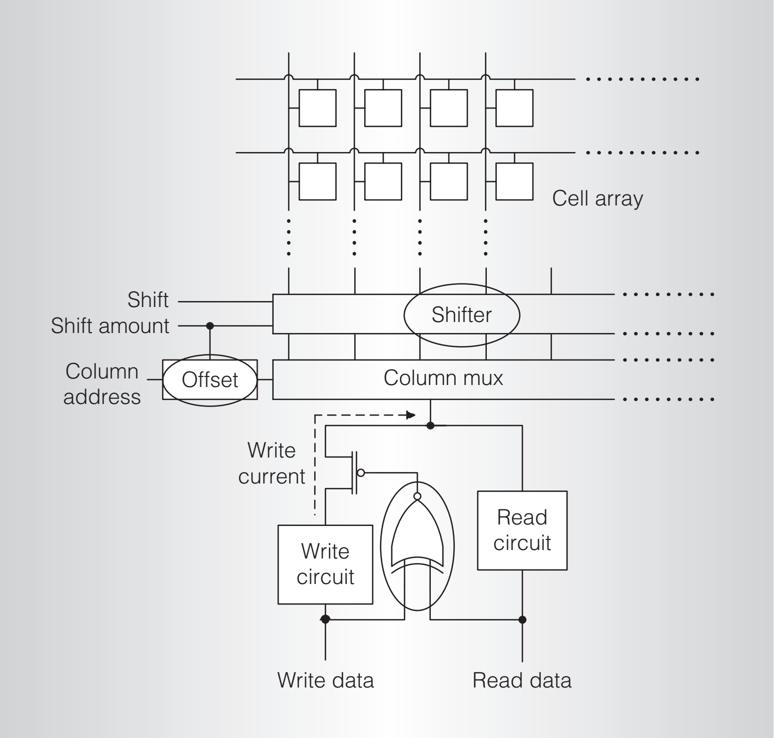

Figure 4. Implementation of redundant bit-write removal and row shifting. Added hardware is circled.

Eliminating redundant bit writes. In conventional memory access, a write updates an entire row of memory cells. However, many of these writes are redundant. Thus, in most cases, writing a cell does not change what is already stored within. In a study with various workloads, 85, 77, and 71 percent of bit writes were redundant for single-level-cell (SLC), multilevel-cell with 2 bits per cell (MLC-2), and multilevel-cell with 4 bits per cell (MLC-4) memories, respectively.

Removing these redundant bit writes can improve the lifetimes of SLC, MLC-2, and MLC-4 PCM-based main memory to 2.1 years, 1.6 years, and 1.4 years, respectively. We implement the scheme by preceding a write with a read and a comparison. After the old value is read, an XNOR gate filters out redundant bit-writes. Because a PCM read is considerably faster than a PCM write, and write operations are typically less latency critical, the negative performance impact of adding a read before a write is relatively small.

Although removing redundant bit writes extends lifetime by approximately a factor of 4, the resulting lifetime is still insufficient for practical purposes. Memory updates tend to exhibit strong locality, such that hot cells fail far sooner than cold cells. Because a memory row or segment's lifetime is determined by the first cell to fail, leveling schemes must distribute writes and avoid creating hot memory regions that impact system lifetime.

Row shifting. After redundant bit writes are removed, the bits that are written most in a row tend to be localized. Hence, a simple shifting scheme can more evenly distribute writes within a row. Experimental studies show that the optimal shift granularity is 1 byte, and the optimal shift interval is 256 writes per row. As Figure 4 shows, the scheme is implemented through an additional row shifter and a shift offset register. On a read access, data is shifted back before being passed to the processor. The delay and energy overhead are counted in the final performance and energy results. Row shifting extends the average lifetimes for SLC, MLC-2, and MLC-4 PCM-based

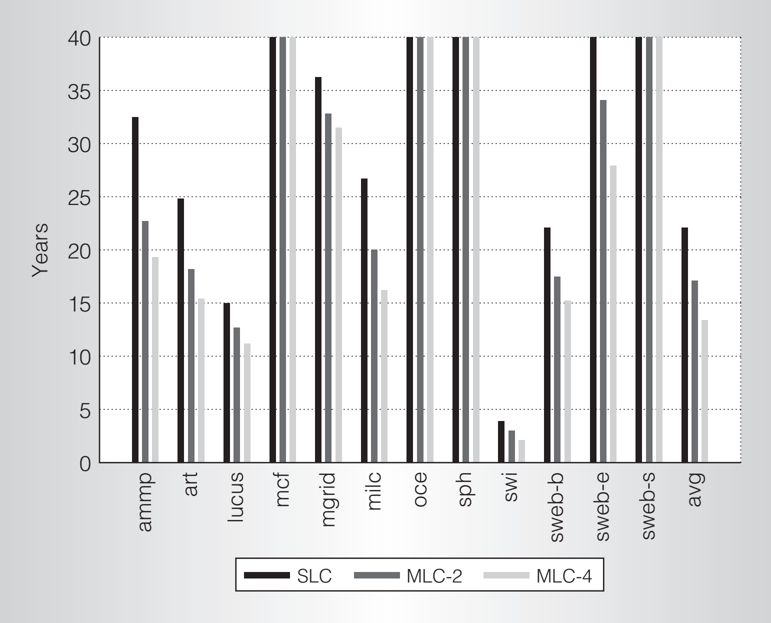

Figure 5. PCM lifetime after eliminating redundant bit writes, row shifting, and segment swapping. [SLC refers to single-level cells. MLC-2 and MLC-4 refer to multilevel cells with 2 and 4 bits per cell, respectively. Some benchmarks (mcf, oce, sph, and sweb-s) exhibit lifetimes greater than 40 years.]

main memories to 5.9 years, 4.4 years, and 3.8 years, respectively.

Segment swapping. The next step considers wear leveling at a coarser granularity: memory segments. Periodically, memory segments of high and low write accesses are swapped. This scheme is implemented in the memory controller, which keeps track of each segment's write counts and a mapping table between the virtual and true segment number. The optimal segment size is 1 Mbyte, and the optimal swap interval is every 2×10^6 writes in each segment. Memory is unavailable in the middle of a swap, which amounts to 0.036 percent performance degradation in the worst case.

Applying the sequence of three techniques[4] extends the average lifetime of PCM-based main memory to 22 years, 17 years, and 13 years, respectively, as Figure 5 shows.

Analyzing energy implications

Because PCM uses an array structure similar to that of DRAM, we use Cacti-D to model energy and delay results for peripheral circuits (interconnections and decoders).[21] We use HSpice simulations to model PCM read operations.[10,22] We derive parameters for PCM write operations from recent results on PCM cells.[3]

Because of its low-leakage and high-density features, we evaluate PCM integrated on top of a multicore architecture using 3D stacking. For baseline DRAM memory, we integrate low-standby-power peripheral logic because of its better energy-delay (ED^2) reduction.[21] For PCM, we use low-operating-power (LOP) peripheral logic because of its low dynamic power, to avoid compounding PCM's already high dynamic energy. Because PCM is nonvolatile, we can power-gate idle memory banks to save leakage. Thus, LOP's higher leakage energy is not a concern for PCM-based main memory.

Energy model. With redundant bit-write removal, the energy of a PCM write is no longer a fixed value. We calculate per-access write energy as follows:

$$E_{\text{pcmwrite}} = E_{\text{fixed}} + E_{\text{read}} + E_{\text{bitchange}}$$

where E_{fixed} is the fixed portion of energy for each PCM write (row selection, decode, XNOR gates, and so on), and E_{read} is the energy to read out the old data for comparison. The variable part, $E_{\text{bitchange}}$, depends on the number of bit writes actually performed:

$$E_{\text{bitchange}} = E_{1\to0}N_{1\to0} + E_{0\to1}N_{0\to1}$$

Performance and energy. Although PCM has slower read and write operations, experimental results show that the performance impact is quite mild, with an average penalty of 5.7 percent. As Figure 6 shows, dynamic energy is reduced by an average of 47 percent relative to DRAM. The savings come from two sources: redundant bit-write removal and PCM's LOP peripheral circuitry, which is particularly power efficient during burst reads. Because of PCM's nonvolatility, we can safely power-gate the idle memory banks without losing data. This, along with PCM's zero cell leakage, results in 70 percent of leakage energy reduction over an already low-leakage DRAM memory.

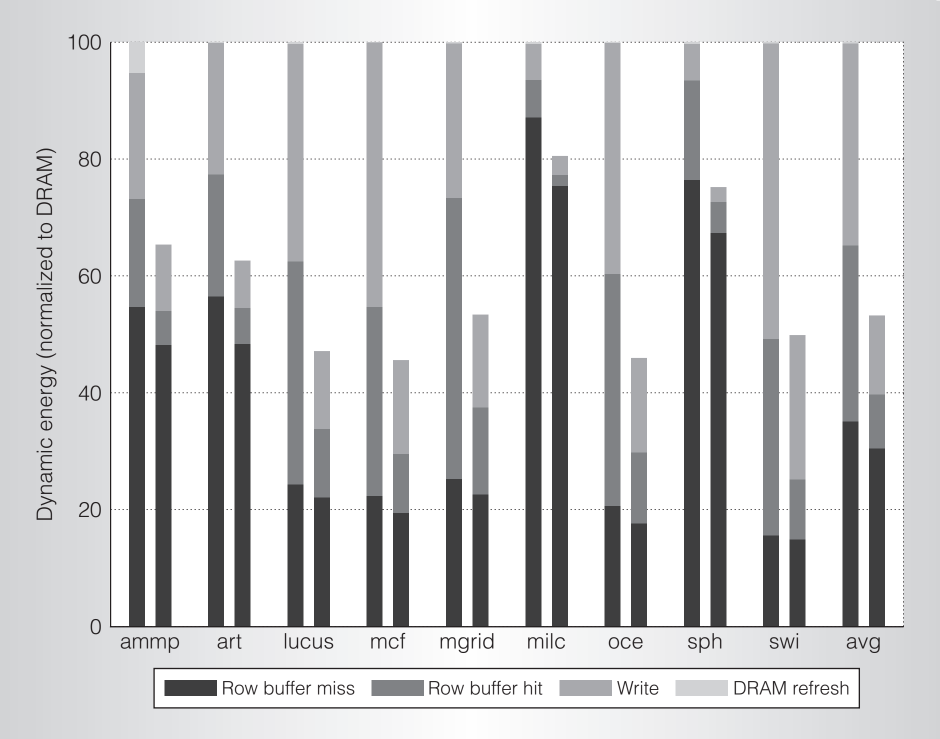

Figure 6. Breakdown of dynamic-energy savings. For each benchmark, the left bar shows the dynamic energy for DRAM, and the right bar shows the dynamic energy for PCM.

Combining dynamic- and leakage-energy savings, we find that the total energy savings is 65 percent, as Figure 7 shows. Because of significant energy savings and mild performance losses, 96 percent of ED^2 reduction is achieved for the ammp benchmark. The average ED^2 reduction for all benchmarks is 60 percent.

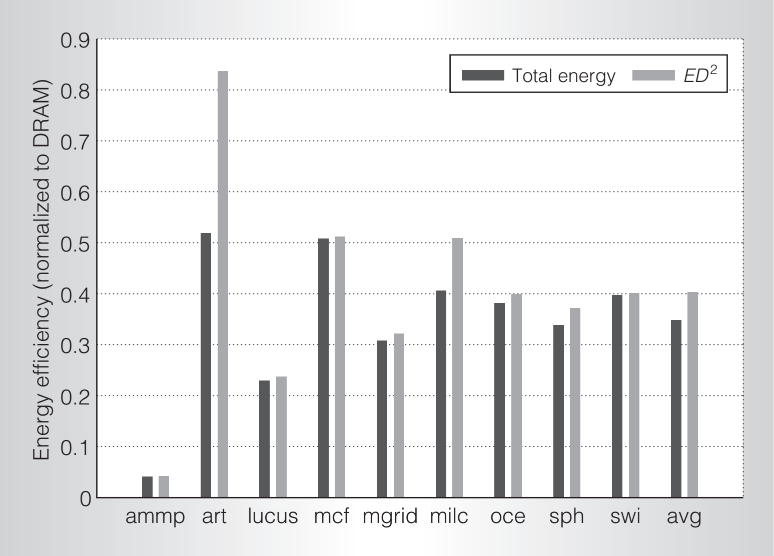

Figure 7. Total energy and energy-delay (ED^2) reduction.

This article has provided a rigorous survey of phase-change technology to drive architectural studies and enhancements. PCM's long latencies, high energy, and finite endurance can be effectively mitigated. Effective buffer organizations, combined with wear reduction and leveling, can make PCM competitive with DRAM at present technology nodes. (Related work also supports this effort.[23,24])

The proposed memory architecture lays the foundation for exploiting PCM scalability and nonvolatility in main memory. Scalability implies lower main-memory energy and greater write endurance. Furthermore, nonvolatile main memories will fundamentally change the landscape of computing. Software designed to exploit the nonvolatility of PCM-based main memories can provide qualitatively new capabilities.

For example, system boot or hibernate could be perceived and instantaneous; application checkpointing could be less expensive;[25] and file systems could provide stronger safety guarantees.[26] Thus, this work is a step toward a fundamentally new memory hierarchy with implications across the hardware-software interface. MICRO

References

1. *Int'l Technology Roadmap for Semiconductors: Process Integration, Devices, and Structures,* Semiconductor Industry Assoc., 2007; http://www.itrs.net/Links/2007ITRS/2007_Chapters/2007_PIDS.pdf.
2. B. Lee et al., ''Architecting Phase Change Memory as a Scalable DRAM Alternative,'' *Proc. 36th Ann. Int'l Symp. Computer Architecture* (ISCA 09), ACM Press, 2009, pp. 2-13.
3. S. Raoux et al., ''Phase-Change Random Access Memory: A Scalable Technology,'' *IBM J. Research and Development,* vol. 52, no. 4, 2008, pp. 465-479.
4. P. Zhou et al., ''A Durable and Energy Efficient Main Memory Using Phase Change Memory Technology,'' *Proc. 36th Ann. Int'l Symp. Computer Architecture* (ISCA 09), ACM Press, 2009, pp. 14-23.
5. Y. Chen et al., ''Ultra-thin Phase-Change Bridge Memory Device Using GeSb,'' *Proc. Int'l Electron Devices Meeting* (IEDM 06), IEEE Press, 2006, pp. 30.3.1-30.3.4.
6. H. Horii et al., ''A Novel Cell Technology Using N-Doped GeSbTe Films for Phase Change RAM,'' *Proc. Symp. VLSI Technology,* IEEE Press, 2003, pp. 177-178.
7. S. Lai, ''Current Status of the Phase Change Memory and Its Future,'' *Proc. Int'l Electron Devices Meeting* (IEDM 03), IEEE Press, 2003, pp. 10.1.1-10.1.4.
8. A. Pirovano et al., ''Scaling Analysis of Phase-Change Memory Technology,'' *Proc. Int'l Electron Devices Meeting* (IEDM 03), IEEE Press, 2003, pp. 29.6.1-29.6.4.
9. F. Bedeschi et al., ''A Multi-level-Cell Bipolar-Selected Phase-Change Memory,'' *Proc. Int'l Solid-State Circuits Conf.* (ISSCC 08), IEEE Press, 2008, pp. 428-429, 625.
10. K.-J. Lee et al., ''A 90 nm 1.8 V 512 Mb Diode-Switch PRAM with 266 MB/s Read Throughput,'' *J. Solid-State Circuits,* vol. 43, no. 1, 2008, pp. 150-162.
11. T. Nirschl et al., ''Write Strategies for 2 and 4-Bit Multi-level Phase-Change Memory,'' *Proc. Int'l Electron Devices Meeting* (IEDM 08), IEEE Press, 2008, pp. 461-464.
12. S. Ahn et al., ''Highly Manufacturable High Density Phase Change Memory of 64Mb and Beyond,'' *Proc. Int'l Electron Devices Meeting* (IEDM 04), IEEE Press, 2004, pp. 907-910.
13. F. Bedeschi et al., ''An 8 Mb Demonstrator for High-Density 1.8 V Phase-Change Memories,'' *Proc. Symp. VLSI Circuits,* 2004, pp. 442-445.
14. H. Oh et al., ''Enhanced Write Performance of a 64 Mb Phase-Change Random Access Memory,'' *Proc. IEEE J. Solid-State Circuits,* vol. 41, no. 1, 2006, pp. 122-126.
15. F. Pellizzer et al., ''A 90 nm Phase Change Memory Technology for Stand-Alone Nonvolatile Memory Applications,'' *Proc. Symp. VLSI Circuits,* IEEE Press, 2006, pp. 122-123.
16. S. Kang et al., ''A 0.1-μm 1.8-V 256-Mb Phase-Change Random Access Memory (PRAM) with 66-MHz Synchronous Burst-Read Operation,'' *IEEE J. Solid-State Circuits,* vol. 42, no. 1, 2007, pp. 210-218.
17. M. Sinha et al., ''High-Performance and Low-Voltage Sense-Amplifier Techniques for Sub-90 nm SRAM,'' *Proc. Int'l Systems-on-Chip Conf.* (SOC 03), 2003, pp. 113-116.
18. V. Aslot and R. Eigenmann, ''Quantitative Performance Analysis of the SPEC OMPM2001 Benchmarks,'' *Scientific Programming,* vol. 11, no. 2, 2003, pp. 105-124.
19. D.H. Bailey et al., *NAS Parallel Benchmarks,* tech. report RNR-94-007, NASA Ames Research Center, 1994.
20. S. Woo et al., ''The SPLASH-2 Programs: Characterization and Methodological Considerations,'' *Proc. 22nd Ann. Int'l Symp. Computer Architecture* (ISCA 1995), ACM Press, 1995, pp. 24-36.
21. S. Thoziyoor et al., ''A Comprehensive Memory Modeling Tool and Its Application to the Design and Analysis of Future Memory Hierarchies,'' *Proc. 35th Int'l Symp. Computer Architecture* (ISCA 08), IEEE CS Press, 2008, pp. 51-62.
22. W.Y. Cho et al., ''A 0.18-μm 3.0-V 64-Mb Nonvolatile Phase-Transition Random Access Memory (PRAM),'' *J. Solid-State Circuits,* vol. 40, no. 1, 2005, pp. 293-300.

23. M.K. Qureshi, V. Srinivasan, and J.A. Rivers, ''Scalable High Performance Main Memory System Using Phase-Change Memory Technology,'' *Proc. 36th Ann. Int'l Symp. Computer Architecture* (ISCA 09), ACM Press, 2009, pp. 24-33.
24. X. Wu et al., ''Hybrid Cache Architecture with Disparate Memory Technologies,'' *Proc. 36th Ann. Int'l Symp. Computer Architecture* (ISCA 09), ACM Press, 2009, pp. 34-45.
25. X. Dong et al., ''Leveraging 3D PCRAM Technologies to Reduce Checkpoint Overhead for Future Exascale Systems,'' *Proc. Int'l Conf. High Performance Computing Networking, Storage, and Analysis* (Supercomputing 09), 2009, article 57.
26. J. Condit et al., ''Better I/O through Byte-Addressable, Persistent Memory,'' *Proc. ACM SIGOPS 22nd Symp. Operating Systems Principles* (SOSP 09), ACM Press, 2009, pp. 133-146.

Benjamin C. Lee is a Computing Innovation Fellow in electrical engineering and a member of the VLSI Research Group at Stanford University. His research focuses on scalable technologies, power-efficient computer architectures, and high-performance applications. Lee has a PhD in computer science from Harvard University.

Ping Zhou is pursuing his PhD in electrical and computer engineering at the University of Pittsburgh. His research interests include new memory technologies, 3D architecture, and chip multiprocessors. Zhou has an MS in computer science from Shanghai Jiao Tong University in China.

Jun Yang is an associate professor in the Electrical and Computer Engineering Department at the University of Pittsburgh. Her research interests include computer architecture, microarchitecture, energy efficiency, and memory hierarchy. Yang has a PhD in computer science from the University of Arizona.

Youtao Zhang is an assistant professor of computer science at the University of Pittsburgh. His research interests include computer architecture, compilers, and system security. Zhang has a PhD in computer science from the University of Arizona.

Bo Zhao is pursuing a PhD in electrical and computer engineering at the University of Pittsburgh. His research interests include VLSI circuits and microarchitectures, memory systems, and modern processor architectures. Zhao has a BS in electrical engineering from Beihang University, Beijing.

Engin Ipek is an assistant professor of computer science and of electrical and computer engineering at the University of Rochester. His research focuses on computer architecture, especially multicore architectures, hardware-software interaction, and high-performance memory systems. Ipek has a PhD in electrical and computer engineering from Cornell University.

Onur Mutlu is an assistant professor of electrical and computer engineering at Carnegie Mellon University. His research focuses on computer architecture and systems. Mutlu has a PhD in electrical and computer engineering from the University of Texas at Austin.

Doug Burger is a principal researcher at Microsoft Research, where he manages the Computer Architecture Group. His research focuses on computer architecture, memory systems, and mobile computing. Burger has a PhD in computer science from the University of Wisconsin, Madison. He is a Distinguished Scientist of the ACM and a Fellow of the IEEE.

Direct questions and comments to Benjamin Lee, Stanford Univ., Electrical Engineering, 353 Serra Mall, Gates Bldg 452, Stanford, CA, 94305; bcclee@stanford.edu.